INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Audio-visual Multi-channel Recognition of Overlapped Speech

Jianwei Yu[1,2], Bo Wu[2], Rongzhi Gu[2], Shi-Xiong Zhang[2], Lianwu Chen[2], Yong Xu[2], Meng Yu[2], Dan Su[2], Dong Yu[2], Xunying Liu[1], Helen Meng[1]

[1]The Chinese University of Hong Kong
[2]Tencent AI Lab

{jwyu,xyliu,hmmmeng}@se.cuhk.edu.hk, {lambowu,auszhang,dyu}@tencent.com

Abstract

Automatic speech recognition (ASR) of overlapped speech remains a highly challenging task to date. To this end, multi-channel microphone array data are widely used in state-of-the-art ASR systems. Motivated by the invariance of visual modality to acoustic signal corruption, this paper presents an audio-visual multi-channel overlapped speech recognition system featuring tightly integrated separation front-end and recognition back-end. A series of audio-visual multi-channel speech separation front-end components based on *TF masking*, *filter&sum* and *mask-based MVDR* beamforming approaches were developed. To reduce the error cost mismatch between the separation and recognition components, they were jointly fine-tuned using the connectionist temporal classification (CTC) loss function, or a multi-task criterion interpolation with scale-invariant signal to noise ratio (Si-SNR) error cost. Experiments suggest that the proposed multi-channel AVSR system outperforms the baseline audio-only ASR system by up to 6.81% (26.83% relative) and 22.22% (56.87% relative) absolute word error rate (WER) reduction on overlapped speech constructed using either simulation or replaying of the lipreading sentence 2 (LRS2) dataset respectively.

Index Terms: Overlapped speech recognition, Speech separation, Audio-visual, Multi-channel

1. Introduction

Despite the rapid progress in automatic speech recognition (ASR) in the past few decades, recognizing overlapped speech remains a highly challenging task. The presence of interfering speakers creates a large mismatch against clean speech, which leads to a significant performance degradation in current ASR systems. To this end, acoustic beamforming techniques integrating sensor data from multiple array channels are usually adopted. These approaches "listen" in the speaker's direction while attenuate the effects of noise distortions and interfering speakers. The desired speaker signal is thereby enhanced. Many state-of-the-art ASR systems have used microphone arrays, often following a traditional speech enhancement based approach. This splits the overall system into two parts: speech separation and speech recognition. The separation components are often realized using conventional beamforming techniques represented either by time domain delay and sum [1, 2] or frequency domain minimum variance distortionless response (MVDR) [3, 4] and generalized eigenvalue (GEV) [5] approaches. The former uses generalized cross entropy with phase transformation and Viterbi search to compute the optimal delay and channel weights, while the latter maximizes the signal to noise ratio (SNR).

The success of deep learning based speech technologies allows microphone array channel integration methods to evolve into a wide range of neural network (NN) based designs. These methods can be roughly classified into three categories, i.e. *TF masking*, *filter&sum* and *mask-based MVDR* or *GEV*. The neural network (NN) based *TF masking* approaches [6, 7] predict spectral time-frequency (TF) masks that specify whether a particular TF bin is dominated by the target speaker or interfering sources to facilitate speech separation. The neural *filter&sum* approaches directly estimate the beamforming filter parameters in either time domain [8–10] or frequency domain [11] before applying these to channel integration to produce the separated output. The more complicated *mask-based MVDR* [12–17] and related *mask-based GEV* [18, 19] approaches compute the power spectral density (PSD) matrices for the target and overlapping speakers using spectral TF masks to obtain the beamforming parameters. The NN based beamforming methods allow tight integration with the recognition back-end to be more conveniently implemented [11, 16, 17, 19, 20]. The use of microphone array based multi-channel inputs can greatly improve the performance of overlapped speech recognition. However, the performance gap between overlapped and non-overlapped speech remains large to date.

Human speech perception is bi-modal in nature [21]. The visual information is inherently invariant to acoustic signal corruption. Therefore, the visual modality can be used to improve the recognition performance on overlapped speech. Previous research has successfully incorporated the visual modality into single-channel overlapped speech separation [22–24] and recognition [25–29]. Recently, audio-visual multi-channel systems designed for speech separation have been proposed in [30, 31]. However, there has been very limited previous research on audio-visual multi-channel recognition of overlapped speech.

In this paper, we proposed an audio-visual multi-channel overlapped speech recognition system featuring tightly integrated separation front-end and recognition back-end. First, a series of audio-visual multi-channel speech separation networks based on *TF masking*, *filter&sum* and *mask-based MVDR* approaches were developed respectively. Second, in order to reduce the error cost mismatch between the separation and recognition components, the two components are jointly fine-tuned using the CTC loss function, or a multi-task criterion interpolation with Si-SNR error cost. Experiments suggest that the proposed audio-visual multi-channel recognition system outperforms the baseline audio-only multi-channel ASR systems by up to 6.81% (26.83% relative) and 22.22% (56.87% relative) absolute WER reduction on overlapped speech constructed using either simulation or replaying of the LRS2 dataset respectively. To the best of our knowledge, this paper is among the first to use audio-visual multi-channel integration for the overlapped speech recognition.

The rest of the paper is organized as follows. Section 2 introduces three neural network based multi-channel integration methods. Section 3 discusses the audio-visual multi-channel speech separation networks. The integration of the separation

This work was done while the author was an intern at Tencent AI Lab.

http://dx.doi.org/10.21437/Interspeech.2020-2346

and recognition components is discussed in section 4. Experimental results are presented in section 5. Section 6 draws the conclusions and discuss possible future directions.

2. Multi-channel Speech Separation

This section introduces the three multi-channel speech separation approaches used in this paper, i.e. *TF masking*, *filter&sum* and *mask-based MVDR*.

2.1. TF masking

The *TF masking* approaches predict spectral TF masks that specify whether a particular TF bin is dominated by the target speaker or the interfering sources to facilitate speech separation. Previous research has shown that the complex mask (CM) [32] outperforms the real-value ratio mask (RM) in speech separation and recognition tasks [20]. Therefore, the CM based TF masking approach is adopted in this work. The complex spectrum of the separated output y_{tf} is computed as follows:

$$y_{tf} = m^s_{tf} * x_{R,tf}, \tag{1}$$

where '$*$' indicates complex multiplication, $x_{R,tf}$ is the reference channel's complex spectrum TF bin of the overlapped speech (without loss of generality, we selects $R = 1$ in this paper), and $m^s_{tf} \in \mathbb{C}$ is the CM of the target speaker. Though the *TF masking* approaches can provide perceptually enhanced sounds, there is a shared belief that the processing artifacts created by the masking are detrimental to the ASR technology [33].

2.2. Filter&sum

The neural *filter&sum* approaches directly estimate the beamforming filter parameters in either time domain [8–10] or frequency domain [11] in a fully-trainable fashion. In this work, we adopt a frequency domain *filter&sum* approach to produce the separated outputs:

$$y_{tf} = \sum_i w_{i,tf} * x_{i,tf}. \tag{2}$$

Where $w_{i,tf}$ is the complex value beamforming parameters corresponding to the ith channel.

2.3. Mask-based MVDR

The more complicated *mask-based MVDR* beamforming approach [12, 34] has demonstrated state-of-the-art performance in noisy and overlapped speech recognition [14–16]. Such approach first uses deep neural networks to estimate the real-value [14–16] or complex [35] TF mask of the target speech m^s_{tf} and other interfering sources m^n_{tf} respectively. The PSD matrices corresponding to each source are then calculated as follows:

$$\begin{aligned}
\mathbf{\Phi}^s_f &= \frac{1}{\sum_{t=1}^T m^s_{tf} * (m^s_{tf})^{\mathsf{H}}} \sum_{t=1}^T (m^s_{tf} * \boldsymbol{x}_{tf})(m^s_{tf} * \boldsymbol{x}_{tf})^{\mathsf{H}}, \\
\mathbf{\Phi}^n_f &= \frac{1}{\sum_{t=1}^T m^n_{tf} * (m^n_{tf})^{\mathsf{H}}} \sum_{t=1}^T (m^n_{tf} * \boldsymbol{x}_{tf})(m^n_{tf} * \boldsymbol{x}_{tf})^{\mathsf{H}},
\end{aligned} \tag{3}$$

where $(\cdot)^{\mathsf{H}}$ denotes the conjugate transpose, $\boldsymbol{x}_{tf} = [x_{1,tf}, .., x_{I,tf}) \in \mathbb{C}^I$ is a complex vector containing the TF bins of all I microphone array channels. $\mathbf{\Phi}^s_f$ and $\mathbf{\Phi}^n_f$ represent the PSD matrices of the target and other interfering sources respectively. The time-invariant beamforming filter parameters $\boldsymbol{w}^s_f$ of the target speech are then obtained by the solution of MVDR beamformer as:

$$\boldsymbol{w}^s_f = \frac{(\mathbf{\Phi}^n_f)^{-1}\mathbf{\Phi}^s_f}{\text{Trace}((\mathbf{\Phi}^n_f)^{-1}\mathbf{\Phi}^s_f)}\boldsymbol{u}, \tag{4}$$

where $\boldsymbol{u} = [1, 0, ..., 0]^T$. Finally, the beamforming filters $\boldsymbol{w}^s_f$ are used to compute the separated spectrum y_{tf} as follows:

$$y_{tf} = (\boldsymbol{w}^s_f)^{\mathsf{H}} \boldsymbol{x}_{tf}. \tag{5}$$

3. Audio-visual Multi-channel Separation

This section presents our audio-visual multi-channel speech separation networks.

Audio inputs: As shown in Figure 1, the complex spectrum of all the microphone array channels are first computed through short-time Fourier transform (STFT). The inter-microphone phase differences (IPDs) [14], which reflect the time difference of arrival (TDOA), are also used as input features:

$$\text{IPD}^{(i,j)}_{tf} = \angle(x_{i,tf}/x_{j,tf}), \tag{6}$$

where $x_{i,tf}$ represents the i-th channel's complex spectrum of the mixed signal at time frame t and frequency bin f, and $\angle(\cdot)$ outputs the angle of the input pair of channel specific TF spectrum. Given the direction of arrival (DOA) of the target speaker, e.g. by tracking the speaker's face from a 180-degree wide-angle camera as shown in Figure 1, a location-guided angle feature (AF) introduced in [13, 30] is adopted to provide the target discriminative information:

$$\begin{aligned}
\text{AF}_{\theta,tf} &= \sum_{m=1}^{M} \langle \boldsymbol{e}^{\text{pd}^{(i,j)}_{\theta,tf}}, \boldsymbol{e}^{\text{IPD}^{(i,j)}_{tf}} \rangle, \\
\text{pd}^{(i,j)}_{\theta,tf} &= 2\pi f f_s d_{ij} \cos(\theta)/(2(F-1)c),
\end{aligned} \tag{7}$$

where $\boldsymbol{e}^{(\cdot)} = [\cos(\cdot), \sin(\cdot)]$, M is the number of selected microphone pairs, $\text{pd}^{(i,j)}_{tf}$ represents the phase delay between ith and jth microphone of a plane wave from direction θ, d_{ij} is the distance between ith and jth microphone, c is the sound velocity, f_s is the sample rate and F is the number of TF bins.

Visual inputs: The visual inputs are shown in the pink part of Figure 1. Considering that the visual modality is invariant to acoustic corruption, this paper leverages the visual modality containing speaker-specific information to improve the estimation of masks or filter parameters. In this work, a LipNet consisting of a 3D convolutional layer and a 18-layer ResNet is used to extract the lip embeddings from the lip region of the target speaker. Such LipNet is first trained on a lipreading task as described in [36]. The lip embeddings extracted by the LipNet are sent into the visual block before being fused with audio modality.

Modality fusion: In this work, we adopt a factorized attention-based modality fusion method proposed in our previous work [30], which has been proven to outperform the concatenation method. This method firstly factorizes the mixed audio into a set of acoustic subspaces, then leverages the target's information from the visual modality to enhance these subspace acoustic embeddings with learnable weights. Please refer to our previous work [30] for details.

The outputs of the fusion layer are sent into the fusion blocks to compute the CM masks or beamforming filter parameters. Figure 1 (a) shows the diagram of the *TF masking* approach, which estimates the CM mask m^s_{tf} of the target speaker. The diagram

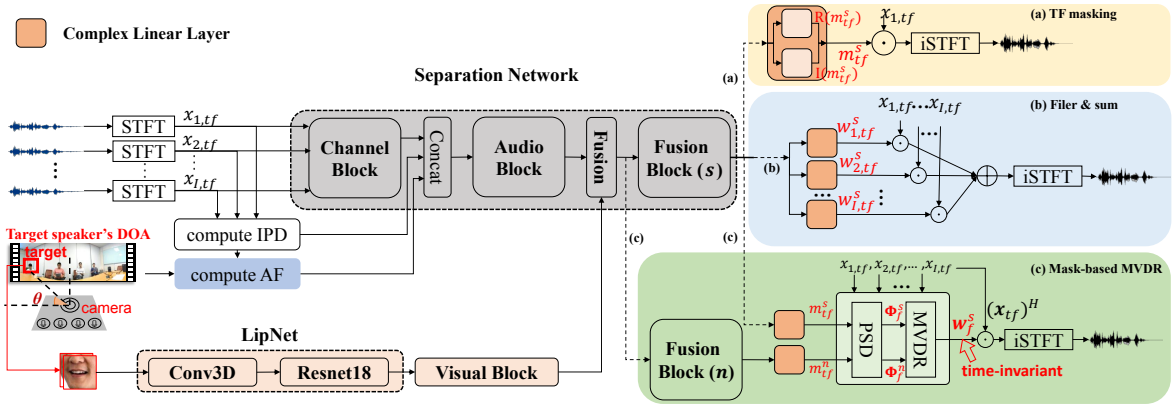

Figure 1: *Illustration of the proposed audio-visual multi-channel speech separation networks, where $x_{i,tf}$ is the complex spectrum of each channel. (a), (b) and (c) represent three options of channel integration approaches: (a) TF masking: m^s_{tf} represents the complex mask of the target speaker, where $\mathsf{R}(m^s_{tf})$ and $\mathsf{I}(m^s_{tf})$ are the real and imaginary part of the mask respectively; (b) Filter&sum: $w^s_{i,tf}$ denotes the beamforming filter parameters of the ith channel; (c) Mask-based MVDR: m^s_{tf} and m^n_{tf} are the complex masks of the target and interfering sources, $\mathbf{\Phi}^s_f$ and $\mathbf{\Phi}^n_f$ are the corresponding PSD matrices, $\boldsymbol{w}^s_f$ is the time-invariant beamforming filter parameters.*

of the *filter&sum* approach is shown in Figure 1 (b), which estimates the beamforming parameters $w_{i,tf}$ of several microphone array channels. The *mask-based MVDR* approach estimates the masks of the target and interfering sources simultaneously before feeding into a MVDR solution layer implementing Eq.(4) and (5), as shown in Figure 1 (c). The Si-SNR loss function is used to train the separation networks. Since dereverberation is beyond the scope of this paper, the reverberant non-overlapped speech signal is used as the supervision, following [9, 30].

4. Integration of Separation & Recognition

Traditionally, the speech separation and recognition components are developed separately and then used in a pipelined fashion [12–15]. However, two issues arise with such approach: 1) the cost function mismatch between separation and recognition components cannot guarantee the separated outputs target to optimal recognition performance; 2) the artifacts created by separation can increase modeling confusion of the recognition component and lead to performance degradation. According to [19, 20, 37, 38], tight integration of the two components with joint fine-tuning can address above two issues.

Recognition network: The architecture of our audio-visual speech recognition (AVSR) network is shown in Figure 2. The lip embeddings extracted from the LipNet is concatenated with the log filter bank acoustic features extracted from the separated waveform. The concatenated features are sent into the convolutional long short-term memory deep neural networks (CLDNN) to generate the frame level mono-phone posteriors. The recognition network is optimized using the CTC loss function.

Integration of separation & recognition: To tightly integrate the separation and recognition components, here we investigate three variants of fine-tuning methods: 1) fine-tuning the recognition system only on the enhanced signals; 2) jointly fine-tuning the separation and recognition components using the CTC cost function; 3) jointly fine-tuning both systems using a multi-task criterion, which interpolates the CTC and Si-SNR cost function:

$$\mathcal{L} = \mathcal{L}_{CTC} + \alpha \mathcal{L}_{Si-SNR}, \tag{8}$$

where α is a manually tuned weight of the Si-SNR loss.

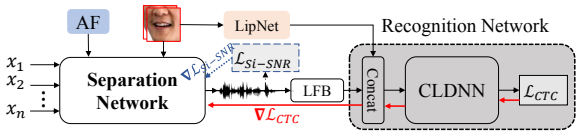

Figure 2: *Joint fine-tuning: $\nabla\mathcal{L}_{CTC}$ and $\nabla\mathcal{L}_{Si-SNR}$ represent the gradients of CTC and SI-SNR loss functions respectively, "LFB" denotes log filter bank acoustic features.*

5. Experiment & Results

5.1. Experiment Setup

Simulated overlapped speech: The multi-channel overlapped speech used in the system development is simulated using the LRS2 dataset [39]. The simulated dataset is split into three subsets with 12.5k, 4.6k and 1.2k utterances respectively for training, validation and evaluation. The details of the simulation process is described in [30].

Replayed overlapped speech: To further evaluate the systems' performance, 1.2k overlapped speech utterances are recorded in a meeting room of the size 10m×5m×3m. To generate overlapped speech, two loudspeakers are used to replay different sentences of the LRS2 test set simultaneously. The structure of the microphone array used during recording is the same as that used in simulation. The target and interfering speakers are located at following directions (15°,30°), (45°,30°), (75°,30°), (105°,30°), (30°,60°), (90°,60°), (120°,60°), (150°,60°) and the distance between the loudspeakers and microphones ranges from 1m to 1.5m. The average overlapping ratio of the replayed overlapped speech is around 80% and SIR is around 1.5dB.

Model architecture: 1) The audio-visual multi-channel separation network is developed based on the time convolutional network (TCN) structure [40]. More details can be found in our previous paper [30]. In the *filter&sum* approach, a series of complex linear layers are used to estimate the filter parameters of each channel using the fusion block's outputs. For all *filter&sum* systems in the remaining part of this paper [1], filter parameters are estimated using only the first and eighth channels. 2) The recognition network starts with four 2-dimensional convolutional layers with channel sizes (64, 64, 128, 128) and kernel size 3x3

[1]Adding more channels up to 15 microphones produces limited improvement for *filter&sum* systems, while increasing the computational cost.

followed by four 1280 hidden units BLSTM layers and one softmax layer. The language model (LM) used in recognition is a 4-gram LM developed on 2.33M words of transcripts of LRS2 training and pretrain set. 3) In multi-task fine-tuning Eq.(8), α is set to 0.1 for *TF masking* approach and 1 for *filter&sum* and *mask-based mvdr* approaches.

Features: 1) For the separation networks, 257-dimensional complex spectrum are used, which are extracted with a 32ms window and 16ms frame rate. The AF and IPDs are extracted between 9 microphone pairs (1,15), (2, 14), (3, 13), (1, 7), (12, 4), (11, 5), (12, 8), (7, 10), (8, 9). These pairs are selected to sample different spacing between microphones [7, 30]. The ground truth direction θ of the target speaker is used during training and evaluation. 2) For the recognition network, 40-dimensional log filter bank features are used, which are extracted using a 40ms window and 10ms frame rate. 3) For visual inputs, we crop the already centered visual frames to 112 by 112 pixels and up-sample them to align with the audio frames via linear interpolation.

5.2. Recognition results on non-overlapped speech

Table 1 presents the WER results of our CLDNN based ASR and AVSR systems on non-overlapped speech in anechoic and simulated reverberant environments. Since we are not aiming for dereverberation in our overlapped speech recognition systems, the WER on reverberant non-overlapped speech (last line) can be viewed as an upper bound for all subsequent experiments on overlapped speech.

Table 1: *Performance of ASR and AVSR systems on echo free and reverberant non-overlapped speech.*

Data	System	WER (%)
Anechoic non-overlapped speech	ASR	11.04
	AVSR	**9.77**
Reverberant non-overlapped speech	ASR	15.33
	AVSR	**13.93**

5.3. Audio-only vs. audio-visual systems

The performance of the audio-only and audio-visual overlapped speech recognition systems trained using simulated overlapped speech is shown in Table 2. All the multi-channel systems (line 3-10) are jointly fine-tuned using the CTC loss function. Several trends can be observed in Table 2. 1) The first block (line 1–2) in Table 2 presents the recognition performance of monaural ASR and AVSR systems without using microphone array and explicit speech separation components. For these very simple systems, simply adding the visual modality in the recognition network can approximately halve the WER, which confirms the findings in our previous research [26]. 2) The second block (line 3–6) of Table 2 shows the results of the multi-channel audio-only systems. "Delay & Sum" means applying frequency domain delay and sum beamforming approach using the steering vector computed by the given array structure and ground truth DOA, as described in [1]. Compared with the monaural ASR system (line 1), the multi-channel speech separation components can significantly improve the systems' performance by up to 49.98% (comparing line 1 & 6 on simu) absolute WER reduction. However, there is a large performance gap (around 13%) between simulated and replayed data using NN based multi-channel audio-only systems (line 4–6). 3) The performance of the proposed audio-visual multi-channel overlapped speech recognition systems is shown in the third block (line 7–10) of Table 2. Comparing the results of audio-visual and audio-only multi-channel systems, it can be seen that leveraging visual modality in both separation and recognition components can reduce the WER ranging from 6.81% (comparing line 6 & 10 on simu) to 28.72% (comparing line 4 & 8 on replay). Moreover, the performance gap between the simulated and replayed overlapped speech is much smaller compared to that on the audio-only systems, which suggests that the proposed audio-visual multi-channel speech recognition systems are more robust.

Table 2: *Performance of audio-only and audio-visual overlapped speech recognition systems using various channel integration methods. The separation and recognition components are jointly fine-tuned using the CTC loss. "AF" denotes angle feature, "raw" denotes raw signal of the first channel.*

	Separation			Recognition	WER(%)	
	method	AF	+visual	+visual	simu	replay
1	raw			✗	75.36	80.55
2	raw			✓	32.06	31.93
3	Delay & sum	✓	✗	✗	49.25	44.34
4	TF masking	✓	✗	✗	33.12	46.75
5	Filter & Sum	✓	✗	✗	30.24	43.83
6	Mask-based MVDR	✓	✗	✗	**25.38**	**39.07**
7	Delay & Sum	✓	✗	✓	25.81	24.46
8	TF masking	✓	✓	✓	19.25	18.03
9	Filter & Sum	✓	✓	✓	**17.21**	19.87
10	Mask-based MVDR	✓	✓	✓	18.57	**16.85**

5.4. Comparison of different fine-tuning approaches

The results of different fine-tuning approaches are listed in Table 3. The first line shows the baseline systems using the CTC loss function to fine-tune the recognition components only while keeping the parameters of the separation components fixed. Jointly fine-tuning the separation and recognition components using the CTC loss function (line 2) can improve the systems' performance by 0.4% to 5.2% WER reduction. The best results are obtained using a multi-task interpolation between the CTC and Si-SNR cost function to fine-tune the entire systems (last line).

Table 3: *Performance of different fine-tuning approaches of audio-visual multi-channel speech recognition systems.*

Fine-tuning			TF masking	Filter&sum	MVDR
Sep.	Recg.	Loss	simu/replay	simu/replay	simu/replay
✗	✓	$\mathcal{L}_{CTC}$	22.9/23.2	19.2/24.1	19.3/17.3
✓	✓	$\mathcal{L}_{CTC}$	19.3/18.0	17.2/19.9	18.6/**16.9**
✓	✓	$\mathcal{L}_{CTC} + \alpha\mathcal{L}_{Si-SNR}$	**18.6/18.0**	**16.1/19.2**	**18.4/16.9**

6. Conclusions & Future Work

This paper presents an audio-visual multi-channel overlapped speech recognition system with tightly integrated separation front-end and recognition back-end. Three multi-channel integration approaches, i.e. *TF masking*, *filter&sum* and *mask-based MVDR* are investigated in the system development. The experiment results suggest that: 1) using visual modality can improve the systems' performance and robustness; 2) jointly fine-tuning the separation and recognition components can tightly integrate the two components for better speech recognition performance. In the future, this work will be extended to: 1) performing separation and dereverberation simultaneously in the separation front-end; 2) applying to more challenging applications, such as the situation when both the visual and audio are degraded; 3) investigating other separation and recognition architectures.

7. Acknowledgements

The authors would like to thank Shansong Liu for the insightful discussion. This research is supported by Hong Kong Research Grants Council General Research Fund No.14200218, Theme Based Research Scheme T45-407/19N and Shun Hing Institute of Advanced Engineering Project No.MMT-p1-19.

8. References

[1] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE assp magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[2] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[3] D. A. Pados and G. N. Karystinos, "An iterative algorithm for the computation of the mvdr filter," *IEEE Transactions On signal processing*, vol. 49, no. 2, pp. 290–300, 2001.

[4] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on audio, speech, and language processing*, vol. 18, no. 2, pp. 260–276, 2009.

[5] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on audio, speech, and language processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[6] F. Bahmaninezhad, J. Wu, R. Gu, S.-X. Zhang, Y. Xu, M. Yu, and D. Yu, "A comprehensive study of speech separation: spectrogram vs waveform separation," *Interspeech*, 2019.

[7] L. Chen, M. Yu, D. Su, and D. Yu, "Multi-band pit and model integration for improved multi-channel speech separation," in *ICASSP*, 2019, pp. 705–709.

[8] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, K. Bacchiani *et al.*, "Multichannel signal processing with deep neural networks for automatic speech recognition," *TASLP*, vol. 25, no. 5, pp. 965–979, 2017.

[9] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "Fasnet: Low-latency adaptive beamforming for multi-microphone audio processing," in *ASRU*, 2019, pp. 260–267.

[10] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *ICASSP*, 2020, pp. 6394–6398.

[11] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *ICASSP*, 2016, pp. 5745–5749.

[12] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved mvdr beamforming using single-channel mask prediction networks." in *Interspeech*, 2016, pp. 1981–1985.

[13] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 558–565.

[14] T. Yoshioka, H. Erdogan, Z. Chen, X. Xiao, and F. Alleva, "Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks," *Interspeech*, 2018.

[15] T. Yoshioka *et al.*, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *ICASSP*, 2018, pp. 5739–5743.

[16] X. Chang, W. Zhang, Y. Qian, J. L. Roux, and S. Watanabe, "Mimo-speech: End-to-end multi-channel multi-speaker speech recognition," *ASRU*, 2019.

[17] X. Chang, W. Zhang, Y. Qian, J. Le Roux, and S. Watanabe, "End-to-end multi-speaker speech recognition with transformer," in *ICASSP*, 2020, pp. 6134–6138.

[18] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *ICASSP*, 2016, pp. 196–200.

[19] J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink, and R. Haeb-Umbach, "Beamnet: End-to-end training of a beamformer-supported multi-channel asr system," in *ICASSP*, 2017, pp. 5325–5329.

[20] Y. Xu, C. Weng, L. Hui, J. Liu, M. Yu, D. Su, and D. Yu, "Joint training of complex ratio mask based beamformer and acoustic model for noise robust asr," in *ICASSP*, 2019, pp. 6745–6749.

[21] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *AAAI*, vol. 33, 2019, pp. 9299–9306.

[22] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," *Interspeech*, 2018.

[23] J. Wu, Y. Xu, S.-X. Zhang, L.-W. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation," *ASRU*, 2019.

[24] T. Afouras, J. S. Chung, and A. Zisserman, "My lips are concealed: Audio-visual speech enhancement through obstructions," *Interspeech*, 2019.

[25] G.-L. Chao, W. Chan, and I. Lane, "Speaker-targeted audio-visual models for speech recognition in cocktail-party environments," *Interspeech*, pp. 2120–2124, 2016.

[26] J. Yu, S.-X. Zhang, J. Wu, S. Ghorbani, B. Wu, S. Kang, S. Liu, X. Liu, H. Meng, and D. Yu, "Audio-visual recognition of overlapped speech for the lrs2 dataset," *ICASSP*, 2020.

[27] B. Garcia, B. Shillingford, H. Liao, O. Siohan, O. d. P. F. Braga, T. Makino, and Y. Assael, "Recurrent neural network transducer for audio-visual speech recognition," *ICASSP*, 2019.

[28] S. Liu, S. Hu, Y. Wang, J. Yu, R. Su, X. Liu, and H. Meng, "Exploiting visual features using bayesian gated neural networks for disordered speech recognition," *Proc. Interspeech 2019*, pp. 4120–4124, 2019.

[29] S. Liu, X. Xie, J. Yu, S. Hu, M. Geng, R. Su, S.-X. Zhang, X. Liu, and H. Meng, "Exploiting cross-domain visual feature generation for disordered speech recognition," *Interspeech*, 2020.

[30] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, "Multi-modal multi-channel target speech separation," *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[31] K. Tan, Y. Xu *et al.*, "Audio-visual speech separation and dereverberation with a two-stage multimodal network," *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[32] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *TASLP*, vol. 24, no. 3, pp. 483–492, 2015.

[33] T. Yoshioka, N. Ito, *et al.*, "The ntt chime-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 436–443.

[34] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel mmse-based framework for speech source separation and noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1913–1928, 2013.

[35] Y. Xu, M. Yu, S.-X. Zhang, L. Chen, C. Weng, J. Liu, and D. Yu, "Neural spatio-temporal beamformer for target speech separation," *Interspeech*, 2020.

[36] T. Afouras, J. S. Chung, and A. Zisserman, "Deep lip reading: a comparison of models and an online application," *Interspeech*, 2018.

[37] T. von Neumann, K. Kinoshita, L. Drude, C. Boeddeker, M. Delcroix, T. Nakatani, and R. Haeb-Umbach, "End-to-end training of time domain audio separation and recognition," in *ICASSP*, 2020, pp. 7004–7008.

[38] M. W. Lam, J. Wang, X. Liu, H. Meng, D. Su, and D. Yu, "Extract, adapt and recognize: an end-to-end neural network for corrupted monaural speech recognition," *Interspeech*, pp. 2778–2782, 2019.

[39] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[40] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *TASLP*, vol. 27, no. 8, pp. 1256–1266, 2019.